AD-A100 562 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
REPRESENTATIONS OF THE SPACE OF DISTRIBUTIONS IN ROBUST ESTIMAT--ETC(U)
MAY 81 D L HALL, B L JOINER DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2219 NL

END
DATE FILMED
7-81
DTIC

LEVEL II

(2)

MRC Technical Summary Report # 2219

REPRESENTATIONS OF THE SPACE OF DISTRIBUTIONS IN ROBUST ESTIMATION OF LOCATION

David L. Hall and Brian L. Joiner

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

May 1981

(Received March 18, 1981)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

REPRESENTATIONS OF THE SPACE OF DISTRIBUTIONS IN ROBUST ESTIMATION OF LOCATION

David L. Hall* and Brian L. Joiner**

Technical Summary Report # 2219

May 1981

ABSTRACT

In many situations it is useful to have a low-dimensional representation of the space of distributions. In this report, one, two and three dimensional representations are given which are of particular relevance to the study of robust estimation of location based on rank estimators. The distances are defined as functions of the asymptotic relative efficiency of the most efficient rank estimator for one distribution when used on data from another distribution. Values of these distance functions are computed for a large number of pairs of distributions and multidimensional scaling is used to find the low-dimensional representations.

AMS (MOS) Subject Classifications: 62F35, 62F12, 62F10, 62G05

Key Words: rank estimators; multidimensional scaling; asymptotic relative efficiency; lambda distributions; contaminated normal distributions; t distributions; adaptive estimation

Work Unit Number 4 - Statistics and Probability

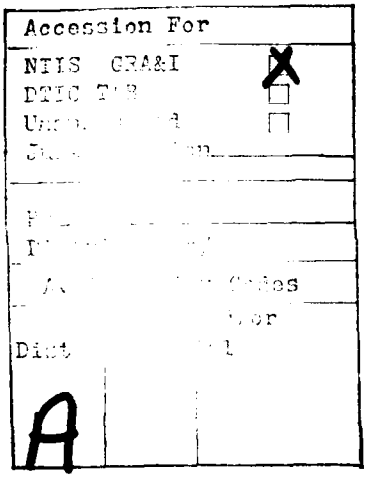
*Associate Manager, Statistics and Materials Safeguards Section, Battelle Northwest Laboratories, Richland, Washington.

**Professor and Director of Statistical Laboratory, Department of Statistics, University of Wisconsin-Madison.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041.

Significance and Explanation

There is considerable interest now as to how one should estimate the center of location of a statistical distribution. Traditionally the sample mean has been used, often in conjunction with outlier rejection rules. However, there are often problems with this procedure. Recent interest has focused on "robust" estimators. This report provides "maps" of an important portion of the space of statistical distributions. These maps are very useful to those studying robust estimators.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

REPRESENTATIONS OF THE SPACE OF DISTRIBUTIONS IN ROBUST ESTIMATION OF LOCATION

David L. Hall* and Brian L. Joiner**

1. Introduction

It is often useful to have a measure of "closeness" among distributions, a way of making more precise such notions as: the normal and logistic distributions are quite similar, whereas the normal and Cauchy are quite different. However, in an important sense, the similarity between distributions is very much a function of the context in which one is working. In some situations, such as variance estimation, the agreement between fourth moments might be critical; in other situations the relative heights of the densities at the medians might be the most important characteristic. In this report we develop "maps" of the space of distributions based on a measure of distance between distributions that is of particular relevance in the problem of robust estimation of location, especially for rank estimators.

The approach used here is intuitively appealing: if two distributions are such that the best estimator for one works quite well on data from the other, the two distributions are in an important sense, quite close. Research in this area apparently begins with the work of Hájek and Šidák (1967). They proposed using as a distance measure a simple function of the asymptotic relative efficiency (ARE) of the corresponding asymptotically most powerful rank tests (amprt). Their measure is $(2(1-\sqrt{ARE}))^{1/2}$. They did not however pursue the idea much further than this definition. Takeuchi, Meisner and Wanderling (1973), hereafter called TMW, presented another related

*Associate Manager, Statistics and Materials Safeguards Section, Battelle Northwest Laboratories, Richland, Washington.

**Professor and Director of Statistical Laboratory, Department of Statistics, University of Wisconsin-Madison.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041.

measure, $\sqrt{1-ARE}$. They computed distances between some pairs of distributions and gave a brief discussion of some of the implications of their distance measure in the context of robust estimation.

These distance measures are not quite as arbitrary as they might at first seem. Recall (Gastwirth, 1966) that a score function for an amprt rank procedure can be viewed as a vector through the origin in Hilbert space and that the square of the cosine of the angle between two such vectors is the ARE of either one applied to data from the other. Then the Hájek and Šidák distance is the chord length of this angle for normalized score functions, and the TMW distance is the sine. We also considered two other distances, the angle itself and its tangent but there seemed to be little practical difference among the four, for present purposes.

We also note that measures of distributional similarity based on ARE are much simpler for rank procedures than for parametric procedures. This results from the reflexivity of the ARE for rank procedures; that is ARE (amprt for F on data from G) = ARE (amprt for G on data from F), a property not possessed by the analogous M and L procedures.

In Hall and Joiner (1980b) the ARE's of rank estimators are computed for a large number of pairs of distributions. Here those efficiencies are converted to distances using the TMW distance (it is easy to prove this is a true metric) and multidimensional scaling (MDS) is used to create low dimensional representations.
The representations depend on the data used and we have chosen to use 45 distributions which are "heavier tailed" than the normal. Distributions with light, uniform-like tails were excluded because our early MDS results indicated that, while the relative locations of the 45 heavier tailed

distributions could be fairly well approximated in a low dimensional space, this was not true of mixtures of heavy tailed and light tailed distributions. We chose to model the heavier tailed distributions since in our experience and from comments in the robustness literature it would appear that heavier tails are more of a problem in practical work.

The representations

Exhibits 1, 2 and 3 give one, two and three dimensional MDS representations of the space spanned by the 45 distributions. All three representations used a regression of the form $y = \theta x$, the standard MDS measure of STRESS and started with four dimensions. Varying these parameters or the choice of distance measure had little effect on the first two dimensions of the fit but did noticeably effect the third dimension. Representations obtained with non-metric MDS were also very similar to the metric ones used here. More details on these alternative solutions are given in Hall (1980).

One useful measure of the adequacy of these or other representations is obtained by considering the fraction of total spread among the points accounted for by the fit. In the representations shown, one dimension accounts for 92% of the spread, two dimensions for 99.3%, three dimensions for 99.8% and four dimensions for 99.9%. Thus for these 45 distributions a fair amount of accuracy is gained by going from one to two dimensions, a bit more by going to three dimensions, but little is gained by going to higher dimensions.

- $\lambda = -1.0$
- Cauchy $(t, \nu=1)$

- Double Exponential

- Logistic-DE $\nu=0.55$
- Student's t $\nu=1.5$
- $\lambda = -0.5$

- CN $\sigma=10$; 10%
- $\lambda = -0.4$

- Student's t $\nu=2$
- Mielke $r=0.2$

- $\lambda = -0.3$
- Logistic-DE $\eta=0.70$

- Mielke $r=0.4$

- $\lambda = -0.2$
- Student's t $\nu=3$
- CN $\sigma=5$; 10%
- Logistic-DE $\eta=0.80$

- $\lambda = -0.1$

Logistic-DE $\eta=0.90$
- Mielke $r=0.8$

CN $\sigma=10$; 5%
- Student's t $\nu=5$

- CN $\sigma=3$; 10%

- Logistic-DE $\eta=0.95$

Logistic
- Logistic-DE $\eta=0.99$
- CN $\sigma=5$; 5%

- Student's t $\nu=10$

Uniform Logistic
- $\lambda = +0.05$
- CN $\sigma=2$; 10%

CN $\sigma=3$; 5%
-

- Mielke $r=1.5$
- CN $\sigma=2$; 5%
- Student's t $\nu=30$

CN $\sigma=5$; 2%
-

- CN $\sigma=10$; 2%
- CN $\sigma=3$; 2%

CN $\sigma=3$; 1%
- CN $\sigma=5$; 1%
- CN $\sigma=2$; 2%
- CN $\sigma=2$; 1%

CN $\sigma=10$; 1%
- Normal

- $\lambda = +0.14$

Exhibit 1

One dimensional representation. This dimension seems at least roughly to correspond to "tail length".

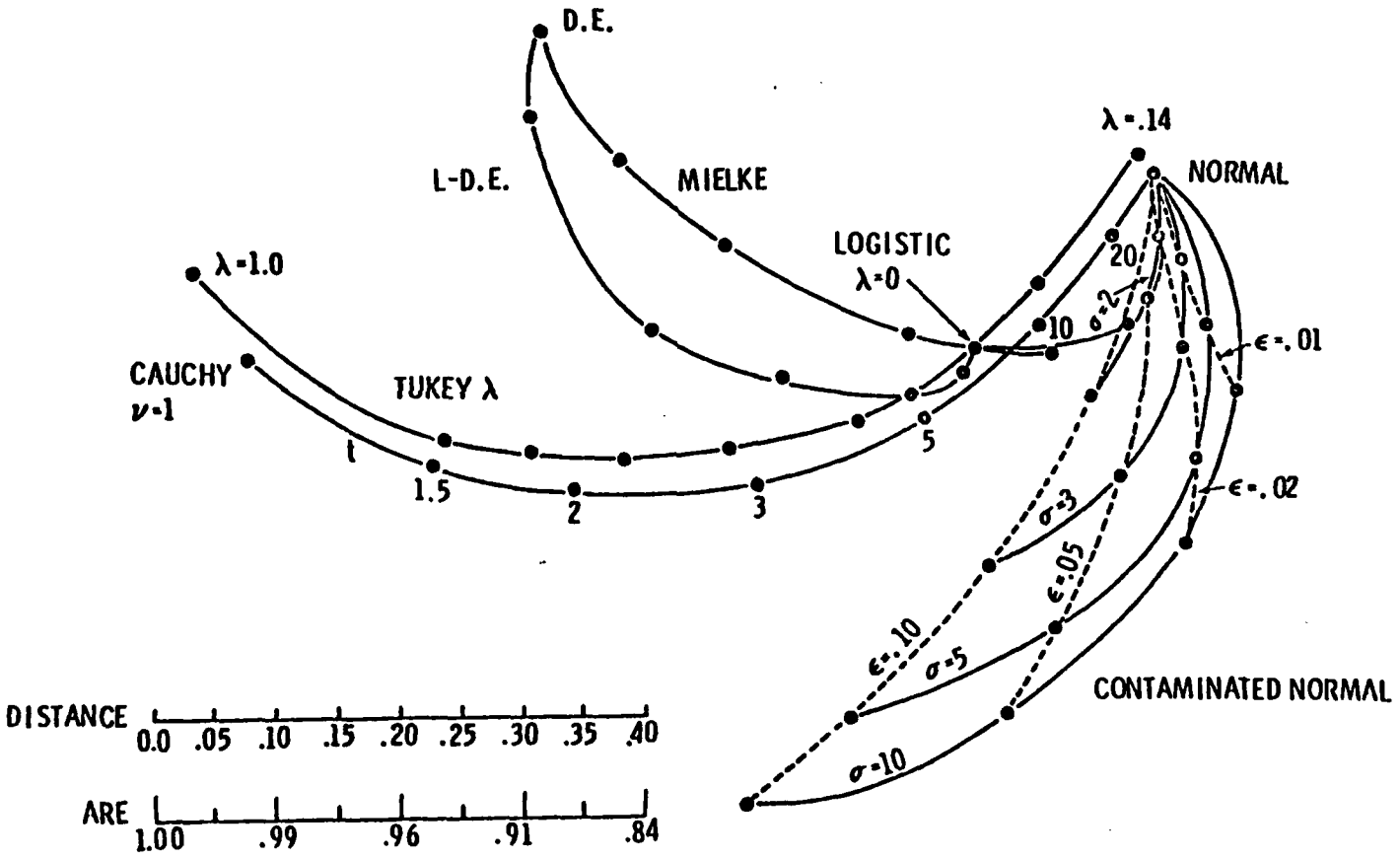


Exhibit 2

Two dimensional representation

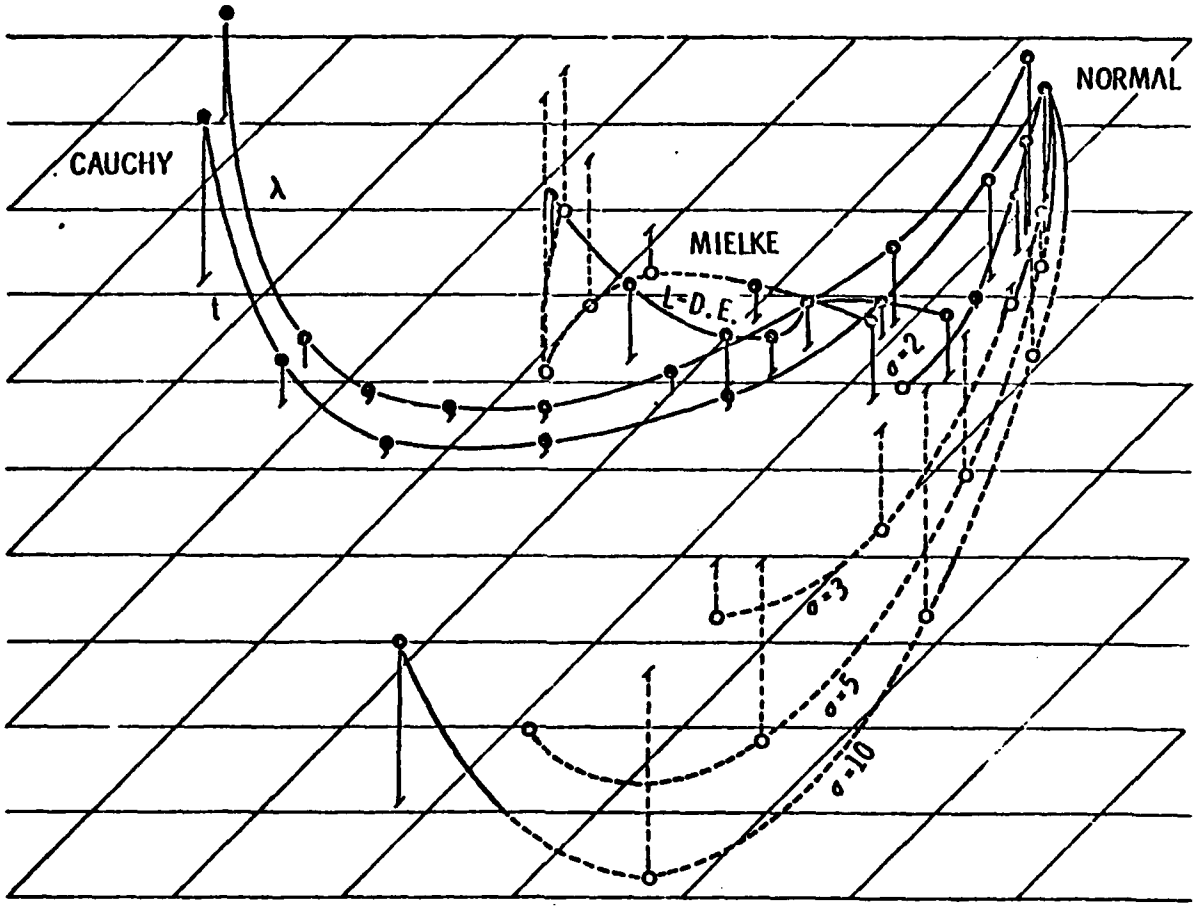


Exhibit 3

Three dimensional representation

Interpretations of Results

Perhaps the most striking aspect of Exhibits 1, 2 and 3 is the much clearer picture afforded by the 2D representation over the more conventional 1D view. The first dimension is by definition the most important and seems to be quite similar to what is ordinarily thought of as "tail weight." Exhibit 2 shows, however, that the second dimension (which does not seem to have any ready interpretation) is almost as important as the first. Having seen this second dimension we find ourselves reluctant to return to any one dimensional representation of this space or use any function that attempts to "order" these distributions. The additional understanding afforded by the third dimension does not seem to be as important as that provided by the first two, thus we find outselves making most use of the 2D figure, referrring to the 3D version only to double check perspectives gained from the 2D representation.

A variety of features are interesting in the 2D and 3D representations. The families flow among relatively smooth curves with the t and λ families in close proximity throughout their range. Good agreement was known between t and λ in other contexts and it was reassuring to see it manifest here. The logistic, a special case of the λ family, is very close to a Student's t with about 7 or 8 degrees of freedom and not too close to the normal. This relates closely to the observation of Mudholkar and Goerge (1978) that the logistic is closer to a t with 9 degrees of freedom than it is to the normal. The t and λ families are relatively one dimensional and fall pretty much along the "tail weight" axis.

The two families that go from the double exponential to the logistic fall along a line that has about a 45° angle with the "tail weight" axis and is almost perpendicular to the contaminated normal range. Thus estimators

or tests based on these two families cannot be expected to do very well on data from t, λ and contaminated normals. The L-DE family in fact corresponds to the family of adaptive rank estimators proposed by Policello and Hettmansperger (1976). Thus it is clear that for contaminated normal data of the sort considered here, the Wilcoxon procedure (corresponding to the logistic distribution) does about as well as the best possible adaptive procedure based on the Policello-Hettmansperger family. For data from the t or λ family, the Wilcoxon could be beaten slightly by the Policello-Hettmansperger family, but clearly it is not the best family for t or λ data.

The median, which corresponds to the double exponential, is clearly a poor choice for data from most all of the distributions considered here, and is especially poor for data from contaminated normals. Thus the median is resistant, in that the value of an estimate is not sensitive to a few serious outliers, but is not efficiency robust for data of the sort considered here. It's efficiency can be quite poor, as low as 65%, corresponding to a wasteage of over one-third of the data.

A natural question that arises is what family would produce estimators that would have relatively high efficiency over the range of distributions considered here. Clearly none of the families we have considered will work. From the general shape of the contaminated normal and t families we are led to conjecture that a family of contaminated t distributions might be rich enough to cover most of this space, except for data near the very peaked double exponential. One might find that the amount of contamination could be fixed at, say, 5% and still provide a contaminated t family rich enough to cover most of the dpace. If so, an adaptive procedure based on a contaminated t family with varying degrees of freedom and varying scale for the contaminent, might suffice.

Selecting a family of estimators that will be rich enough to be useful is thus one obvious use of these representations. Another important use is to help select representative distributions to use to generate the data for Monte Carlo and other studies of the properties of robust estimators.

References

Hájek, J. and Šidák, Z. (1967). Theory of Rank Tests. Academic Press, New York.

Hall, David Lynn (1980). "Aspects of efficiency-robust estimation of location". Unpublished Ph.D. Thesis, University of Wisconsin, Madison.

Hall, David L. and Joiner, B.L. (1980b), "Asymptotic relative efficiencies of R-estimators: Some numerical and analytic results". (Unpublished manuscript.)

Mudholkar, Govind S. and George, E.O. (1978). "A remark on the shape of the logistic distribution", Biometrika, 65, 667-688.

Takeuchi, K., Meisner, M., and Wanderling, J. (1973). "Asymptotic efficiencies of estimators of location: a computational study". J. Statist. Comput. Simul., 2, 375-390.

DL4/BLJ/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2219	2. GOVT ACCESSION NO. AD-A100 562	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Representations of the Space of Distributions in Robust Estimation of Location		5. TYPE OF REPORT & PERIOD COVERED Summary Report, no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) David L. Hall and Brian L. Joiner		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE May 1981 13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

rank estimators; multidimensional scaling; asymptotic relative efficiency; lambda distributions; contaminated normal distributions; t distributions; adaptive estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In many situations it is useful to have a low-dimensional representation of the space of distributions. In this report, one, two and three dimensional representations are given which are of particular relevance to the study of robust estimation of location based on rank estimators. The distances are defined as functions of the asymptotic relative efficiency of the most efficient rank estimator for one distribution when used on data from another distribution. Values of these distance functions are computed for a large number of pairs of distributions and multidimensional scaling is used to find the low-dimensional representations.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


DA
FILM
7